

Configuring cloud data access 2

## Configuring Cloud Data Access

**Date of Publish:** 2018-09-14



<http://docs.hortonworks.com>

# Contents

<b>Configuring access to Amazon S3.....</b>	<b>3</b>
Create an IAM role for S3 access.....	3
Configure access to S3.....	5
Test access from HDP to S3.....	5
Configure S3 storage locations.....	6
<b>Configuring access to ADLS.....</b>	<b>7</b>
Prerequisites.....	7
Configure access to ADLS.....	7
Test access to ADLS.....	8
Configure ADLS storage locations.....	8
<b>Configuring access to WASB.....</b>	<b>9</b>
Prerequisites.....	9
Configure access to WASB.....	9
Test access to WASB.....	10
Configure WASB storage locations.....	11
<b>Configuring access to GCS.....</b>	<b>11</b>
Prerequisites.....	12
Configure access to GCS.....	13
Test access to GCS.....	14
Configuring GCS storage locations.....	14

## Configuring access to Amazon S3

Amazon S3 is not supported as default file system, but access to data in S3 is possible via the s3a connector. Use these steps to configure access from your cluster to Amazon S3.

These steps assume that you are using an HDP version that supports the s3a cloud storage connector (HDP 2.6.1 or newer).

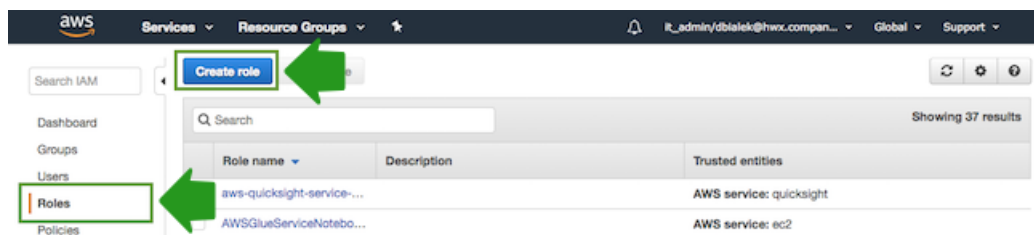
### Create an IAM role for S3 access

In order to configure access from your cluster to Amazon S3, you must have an existing IAM role which determines what actions can be performed on which S3 buckets.

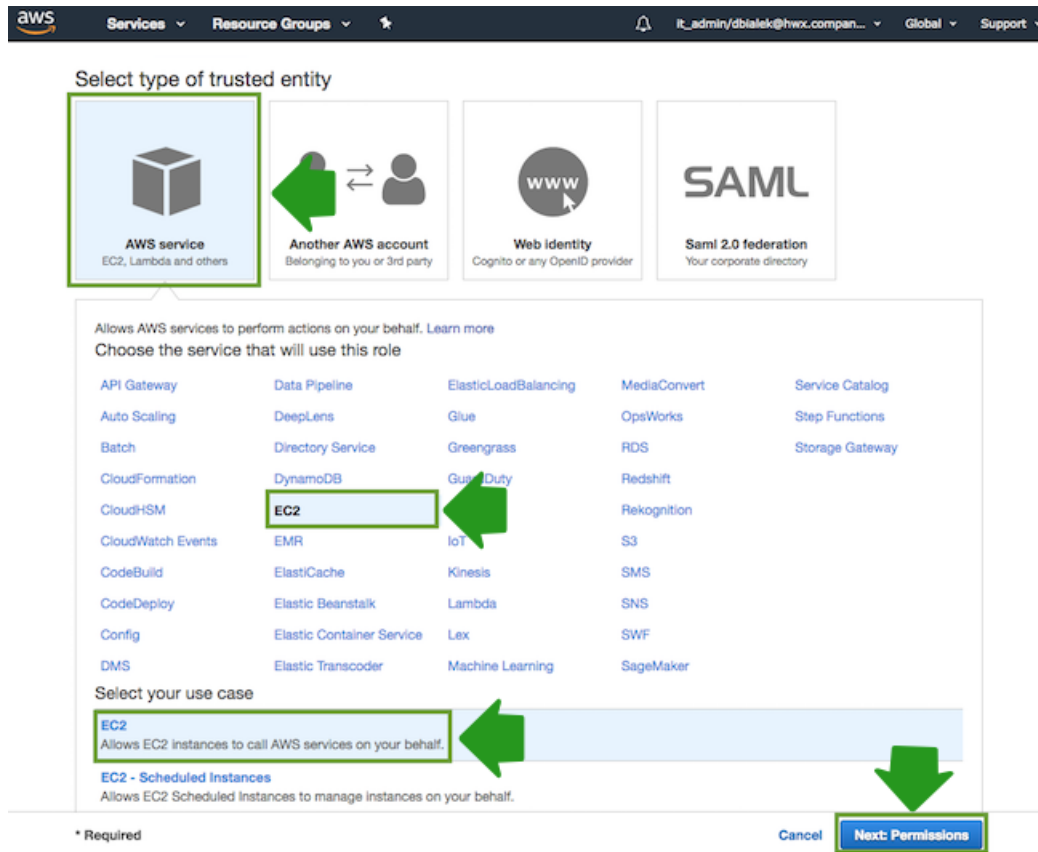
If you already have an IAM role, skip to the next step. If you do not have an existing IAM role, use the following instructions to create one.

Steps

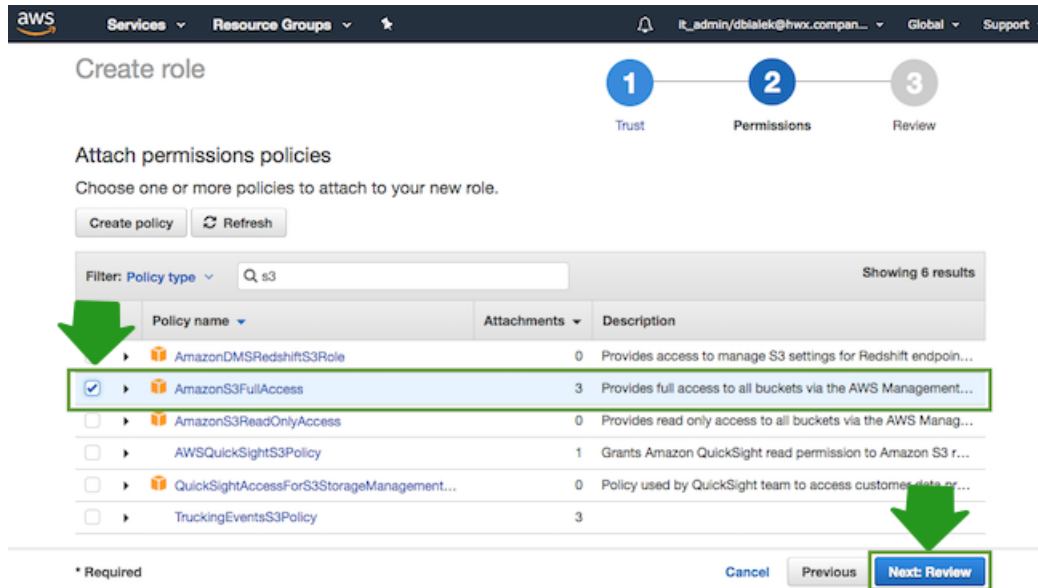
1. Navigate to the IAM console > Roles and click Create Role.



2. In the “Create Role” wizard, select AWS service role type and then select EC2 service and EC2 use case.



3. When done, click Next: Permissions to navigate to the next page in the wizard.
4. Select an existing S3 access policy or click Create policy to define a new policy. If you are just getting started, you can select a built-in policy called “AmazonS3FullAccess”, which provides full access to S3 buckets that are part of your account:



5. When done attaching the policy, click Next: Review.
6. In the Roles name field, enter a name for the role that you are creating:

**Create role**

1 Trust    2 Permissions    3 Review

**Review**

Provide the required information below and review this role before you create it.

**Role name\*** s3access-role  
Maximum 64 characters. Use alphanumeric and '+,=,@,-' characters.

**Role description** Allows EC2 instances to call AWS services on your behalf.  
Maximum 1000 characters. Use alphanumeric and '+,=,@,-' characters.

**Trusted entities** AWS service: ec2.amazonaws.com

**Policies** AmazonS3FullAccess

\* Required    Cancel    Previous    **Create role**

7. Click Create role to finish the role creation process.

## Configure access to S3

Access to data in S3 from your cluster VMs can be automatically configured by attaching an existing instance profile allowing access to S3.

### Prerequisites

You or your AWS admin must create an IAM role with an S3 access policy which can be used by cluster instances to access one or more S3 buckets. Refer to [Create an IAM role for S3 access](#).

### Steps

1. On the Cloud Storage page in the advanced cluster wizard view, select Use existing instance profile.
2. Select an existing IAM role created in step 1.
3. During the cluster creation process, Cloudbreak assigns the IAM role and its associated permissions to the EC2 instances that are part of the cluster so that applications running on these instances can use the role to access S3. Once your cluster is in the running state, you will be able to access the S3 bucket from the cluster nodes.

### Related Information

[Create an IAM role for S3 access](#)

[Using an IAM Role to Grant Permissions to Applications Running on Amazon EC2 Instances \(AWS\)](#)

[Creating a cluster](#)

## Test access from HDP to S3

To test access to S3 from HDP, SSH to a cluster node and run a few hadoop fs shell commands against your existing S3 bucket.

To tests access to S3 from HDP, SSH to a cluster node and run a few hadoop fs shell commands against your existing S3 bucket.

Amazon S3 access path syntax is:

```
s3a://bucket/dir/file
```

For example, to access a file called “mytestfile” in a directory called “mytestdir”, which is stored in a bucket called “mytestbucket”, the URL is:

```
s3a://mytestbucket/mytestdir/mytestfile
```

The following FileSystem shell commands demonstrate access to a bucket named “mytestbucket”:

```
hadoop fs -ls s3a://mytestbucket/  
hadoop fs -mkdir s3a://mytestbucket/testDir  
hadoop fs -put testFile s3a://mytestbucket/testFile  
hadoop fs -cat s3a://mytestbucket/testFile  
test file content
```

For more information about configuring the S3 connector for HDP and working with data stored on S3, refer to [Cloud Data Access HDP](#) documentation.

### Related Information

[Cloud data access \(HDP\)](#)

## Configure S3 storage locations

After configuring access to S3 via instance profile, you can optionally use an S3 bucket as a base storage location; this storage location is mainly for the Hive Warehouse Directory (used for storing the table data for managed tables).

### Prerequisites

- You must have an existing bucket. For instructions on how to create a bucket on S3, refer to AWS documentation.
- The instance profile that you configured must allow access to the bucket.

### Steps

1. When creating a cluster, on the Cloud Storage page in the advanced cluster wizard view, select Use existing instance profile and select the instance profile to use, as described in the documentation for configuring access to S3.
2. Under Storage Locations, enable Configure Storage Locations by clicking the



button.

3. Provide your existing bucket name under Base Storage Location.



#### Note:

Make sure that the bucket already exists within the account.

4. Under Path for Hive Warehouse Directory property (hive.metastore.warehouse.dir), Cloudbreak automatically suggests a location within the bucket. For example, if the bucket that you specified is my-test-bucket then the suggested location will be my-test-bucket/apps/hive/warehouse. You may optionally update this path or select Do not configure.



#### Note:

This directory structure will be created in your specified bucket upon the first activity in Hive.

### Related Information

[Configure access to S3](#)

[Create a bucket \(AWS\)](#)

## Configuring access to ADLS

Hortonworks Data Platform (HDP) supports reading and writing files stored in an Azure Data Lake Storage (ADLS) account.

Azure Data Lake Store (ADLS) is an enterprise-wide hyper-scale repository for big data analytic workloads. ADLS is not supported as default file system, but access to data in ADLS is possible via the adl connector.

These steps assume that you are using an HDP version that supports the adl cloud storage connector (HDP 2.6.1 or newer).

### Related Information

[Overview of Azure Data Lake Store \(Azure\)](#)

## Prerequisites

If you would like to use ADLS to store your data, you must enable Azure subscription for Data Lake Store, and then create an Azure Data Lake Store storage account.

For instructions on how to get started with ADLS, refer to Azure documentation.

### Related Information

[Get started with Azure Data Lake Store using the Azure portal \(Azure\)](#)

## Configure access to ADLS

To configure authentication with ADLS using the client credential, you must register a new application with Active Directory service and then give your application access to your ADL account. After you've performed these steps, you can provide your application's information when creating a cluster.

### Prerequisites

Register an application and add it to the ADLS account, as described in Step 1 and Step 2 of the HCC article [How to configure authentication with ADLS](#).



### Note:

Do not perform the Step 3 described in this article. Cloudbreak automates this step.

### Steps

1. In Cloudbreak web UI, on the advanced Cloud Storage page of the create a cluster wizard, select Use existing ADLS storage.
2. Provide the following parameters for your registered application:
  - ADLS Account Name: This is the ADL account that your application was assigned to.
  - Application ID: You can find it in your application's settings.
  - Key: This is the key that you generated for your application. If you did not copy the it, you must create a new key from the Keys page in your application's settings.
3. Once your cluster is in the running state, you will be able to access the Azure blob storage account from the cluster nodes.

### Related Information

[How to configure authentication with ADLS \(HCC\)](#)

[Creating a cluster](#)

## Test access to ADLS

To test access to ADLS, SSH to a cluster node and run a few `hadoop fs` shell commands against your existing ADLS account.

ADLS access path syntax is:

```
adl://account_name.azuredatalakestore.net/dir/file
```

For example, the following Hadoop FileSystem shell commands demonstrate access to a storage account named “myaccount”:

```
hadoop fs -mkdir adl://myaccount.azuredatalakestore.net/testdir
hadoop fs -put testfile adl://myaccount.azuredatalakestore.net/testdir/
testfile
```

To use `DistCp` against ADLS, use the following syntax:

```
hadoop distcp
  [-D hadoop.security.credential.provider.path=localjceks://file/home/
  user/adls.jceks]
  hdfs://namenode_hostname:9001/user/foo/007020615
  adl://myaccount.azuredatalakestore.net/testDir/
```

For more information about configuring the ADLS connector and working with data stored in ADLS, refer to [Cloud Data Access HDP](#) documentation.

### Related Information

[Cloud data access \(HDP\)](#)

## Configure ADLS storage locations

After configuring access to ADLS, you can optionally use that ADLS storage account as a base storage location; this storage location is mainly for the Hive Warehouse Directory (used for storing the table data for managed tables).

Steps

1. When creating a cluster, on the Cloud Storage page in the advanced cluster wizard view, select Use existing ADLS storage and enter information related to your ADLS account, as described in the instructions for configuring access to ADLS.
2. Under Storage Locations, enable Configure Storage Locations by clicking the



button.

3. Provide your existing directory name under Base Storage Location.



**Note:**

Make sure that the bucket already exists within the account.

4. Under Path for Hive Warehouse Directory property (`hive.metastore.warehouse.dir`), Cloudbreak automatically suggests a location within the directory. For example, if the directory that you specified is `my-test-dir` then the suggested location will be `my-test-adls-account.azuredatalakestore.net/my-test-dir/apps/hive/warehouse`. You may optionally update this path or select Do not configure.



**Note:**

This directory structure will be created in your specified container upon the first activity in Hive.



### Related Information

[Configure access to ADLS](#)

## Configuring access to WASB

Hortonworks Data Platform (HDP) supports reading and writing block blobs and page blobs from and to Windows Azure Storage Blob (WASB) object store.

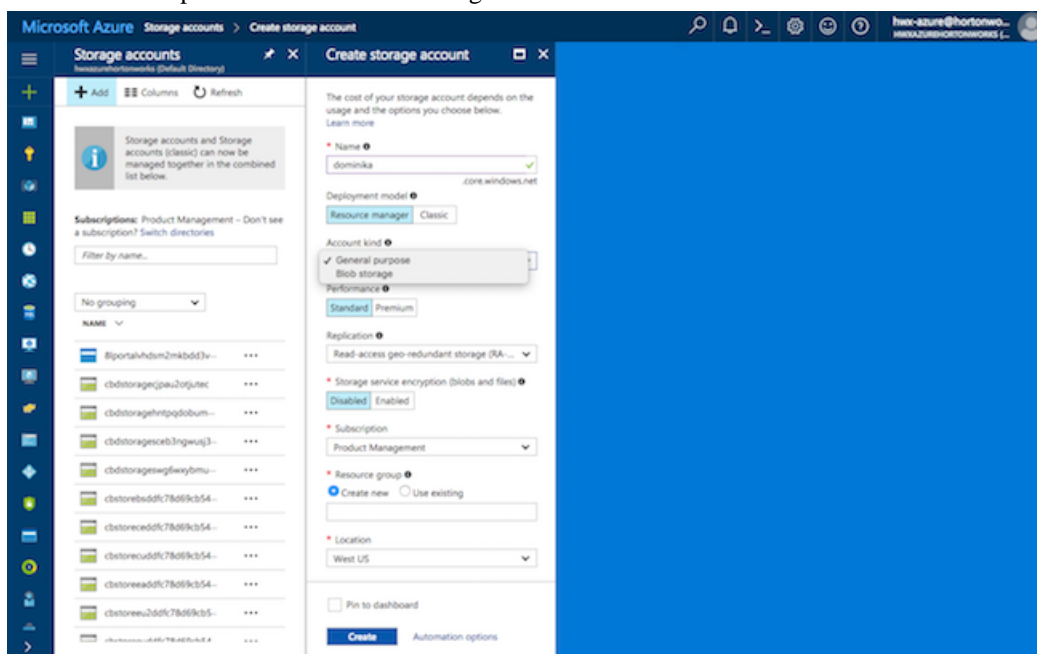
Windows Azure Storage Blob (WASB) is an object store service available on Azure. WASB is not supported as default file system, but access to data in WASB is possible via the wasb connector.

These steps assume that you are using an HDP version that supports the wasb cloud storage connector (HDP 2.6.1 or newer).

### Prerequisites

If you want to use Windows Azure Storage Blob to store your data, you must enable Azure subscription for Blob Storage and then create a storage account.

Here is an example of how to create a storage account:



For detailed instructions, refer to [Azure documentation](#).

### Related Information

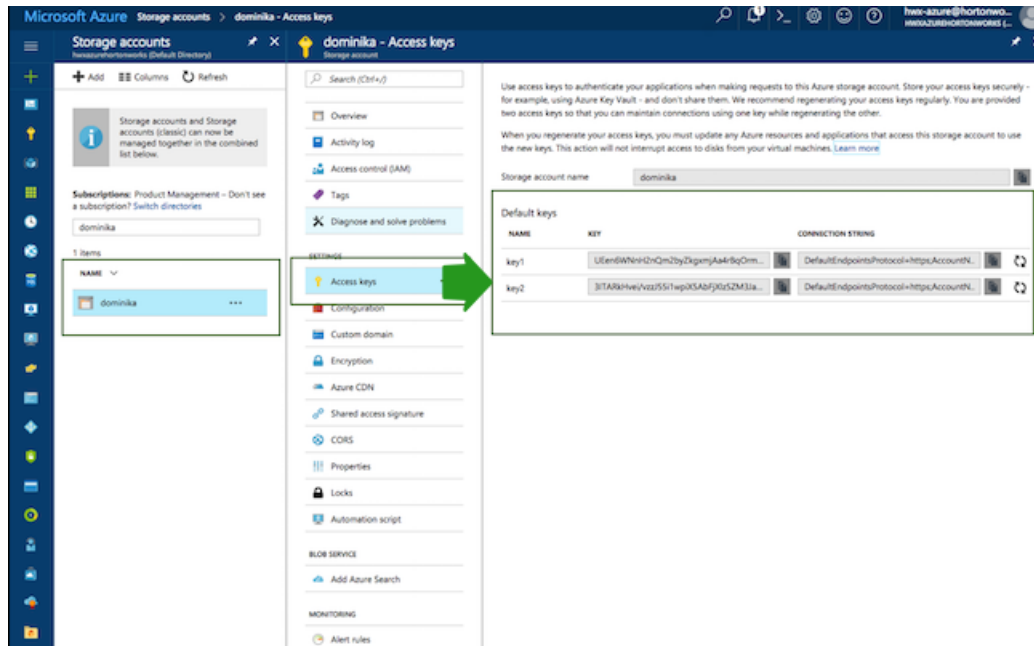
[Create a storage account \(Azure\)](#)

## Configure access to WASB

In order to access data stored in your Azure blob storage account, you must obtain your access key from your storage account settings, and then provide the storage account name and its corresponding access key when creating a cluster.

Steps

1. Obtain your storage account name and storage access key from the Access keys page in your storage account settings:



2. In Cloudbreak web UI, on the advanced Cloud Storage page of the create a cluster wizard:
  - a. Select Use existing WASB storage.
  - b. Provide the Storage Account Name and Access Key parameters obtained in the previous step.
3. Once your cluster is in the running state, you will be able to access ADLS from the cluster nodes.

### Related Information

[Creating a cluster](#)

## Test access to WASB

To test access to WASB, SSH to a cluster node and run a few hadoop fs shell commands against your existing WASB account.

WASB access path syntax is:

```
wasb://container_name@storage_account_name.blob.core.windows.net/dir/file
```

For example, to access a file called "testfile" located in a directory called "testdir", stored in the container called "testcontainer" on the account called "hortonworks", the URL is:

```
wasb://testcontainer@hortonworks.blob.core.windows.net/testdir/testfile
```

You can also use "wasbs" prefix to utilize SSL-encrypted HTTPS access:

```
wasbs://<container_name>@<storage_account_name>.blob.core.windows.net/dir/file</pre>
```

The following Hadoop FileSystem shell commands demonstrate access to a storage account named "myaccount" and a container named "mycontainer":

```
hadoop fs -ls wasb://mycontainer@myaccount.blob.core.windows.net/
hadoop fs -mkdir wasb://mycontainer@myaccount.blob.core.windows.net/testDir
```

```
hadoop fs -put testFile wasb://mycontainer@myaccount.blob.core.windows.net/  
testDir/testFile  
  
hadoop fs -cat wasb://mycontainer@myaccount.blob.core.windows.net/testDir/  
testFile  
test file content
```

For more information about configuring the ADLS connector and working with data stored in ADLS, refer to [Cloud Data Access HDP](#) documentation.

### Related Information

[Cloud data access \(HDP\)](#)

## Configure WASB storage locations

After configuring access to WASB, you can optionally use that WASB storage account as a base storage location; this storage location is mainly for the Hive Warehouse Directory (used for storing the table data for managed tables).

1. When creating a cluster, on the Cloud Storage page in the advanced cluster wizard view, select Use existing WASB storage and enter information related to your WASB account, as described in the instructions for configuring access to WASB.
2. Under Storage Locations, enable Configure Storage Locations by clicking the



button.

3. Provide your existing container name under Existing Base Storage Location.



#### Note:

Make sure that the container already exists within your storage account.

4. Under Path for Hive Warehouse Directory property (hive.metastore.warehouse.dir), Cloudbreak automatically suggests a location within the container. For example, if the directory that you specified is my-test-container then the suggested location will be my-test-container@my-wasb-account.blob.core.windows.net/apps/hive/warehouse. You may optionally update this path or select Do not configure.



#### Note:

This directory structure will be created in your specified container upon the first activity in Hive.

### Related Information

[Configure access to WASB](#)

## Configuring access to GCS

Google Cloud Storage (GCS) is not supported as default file system, but access to data in GCS is possible via the gs connector. Use these steps to configure access from your cluster to GCS.

These steps assume that you are using an HDP version that supports the gs cloud storage connector (HDP 2.6.5 introduced this feature as a TP).

### Related Information

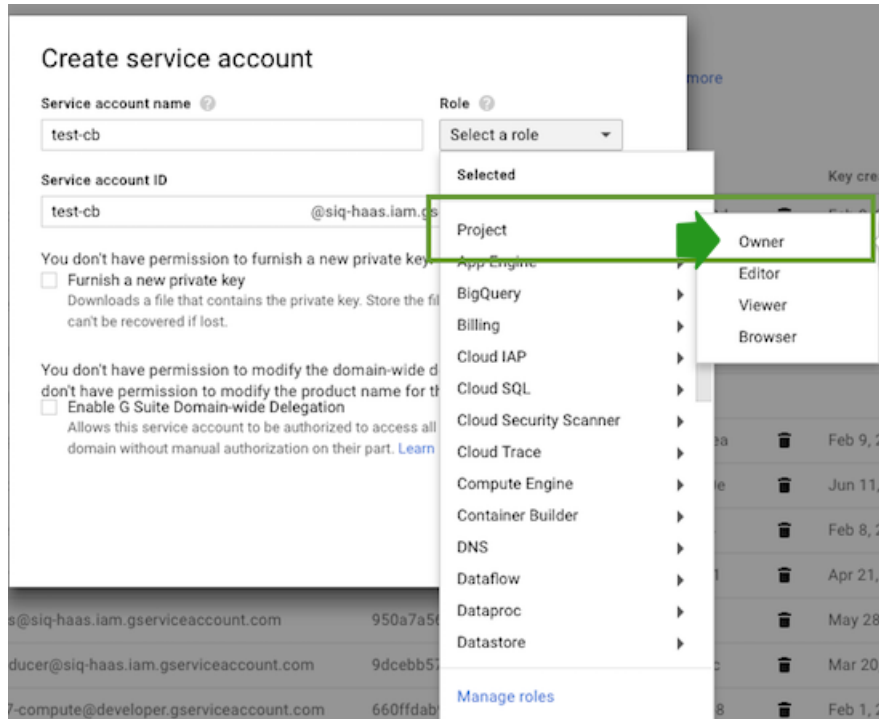
[Cloud Storage Documentation \(Google\)](#)

## Prerequisites

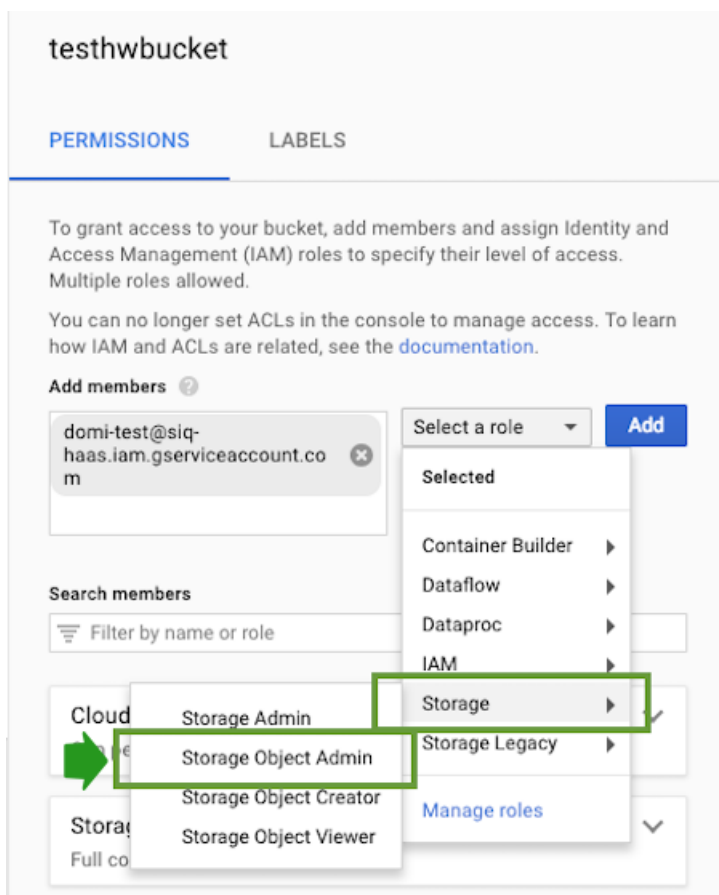
Access to Google Cloud Storage is via a service account. The service account used for access to GCS must meet the minimum requirements.

The service account that you provide to Cloudbreak for GCS data access must have the following permissions:

- The service account must have the project-wide Owner role:



- The service account must have the Storage Object Admin role for the bucket that you are planning to use. You can set this in the bucket's permissions settings:



For more information on service accounts and instructions how to create one, refer to [Google documentation](#).

#### Related Information

[Understanding service accounts \(Google\)](#)

[Creating and managing service accounts \(Google\)](#)

## Configure access to GCS

Access to Google Cloud Storage can be configured separately for each cluster by providing the service account email address.

#### Steps

1. On the Cloud Storage page in the advanced cluster wizard view, select Use existing GSC storage.
2. Under Service Account Email Address provide the email of the service account that you created as a prerequisite.
3. Once your cluster is in the running state, you should be able to access buckets that the configured storage account has access to.

#### Related Information

[Prerequisites](#)

[Creating a cluster](#)

## Test access to GCS

Test access to the Google Cloud Storage bucket by running a few commands from any cluster node. For example, you can use the command listed below (replace “mytestbucket” with the name of your bucket):

```
hadoop fs -ls gs://my-test-bucket/
```

For more information about configuring the gs connector for HDP and working with data stored on GCS, refer to [Cloud Data Access HDP documentation](#).

### Related Information

[Cloud data access \(HDP\)](#)

## Configuring GCS storage locations

After configuring access to GCS via service account, you can optionally use a GCS bucket as a base storage location; this storage location is mainly for the Hive Warehouse Directory (used for storing the table data for managed tables).

### Prerequisites

- You must have an existing bucket. For instructions on how to create a bucket on GCS, refer to [GCP documentation](#).
- The service account that you configured must allow access to the bucket.

### Steps

1. When creating a cluster, on the Cloud Storage page in the advanced cluster wizard view, select Use existing GCS storage and enter information related to your GCS account, as described in the instructions for configuring access to GCP.
2. Under Storage Locations, enable Configure Storage Locations by clicking the



button.

3. Provide your existing directory name under Base Storage Location.



#### Note:

Make sure that the bucket already exists within the account.

4. Under Path for Hive Warehouse Directory property (hive.metastore.warehouse.dir), Cloudbreak automatically suggests a location within the bucket. For example, if the bucket that you specified is my-test-bucket then the suggested location will be my-test-bucket/apps/hive/warehouse. You may optionally update this path or select Do not configure.



#### Note:

This directory structure will be created in your specified bucket upon the first activity in Hive.

### Related Information

[Configure access to GCS](#)

[Creating buckets \(Google\)](#)